

Wojciech Hyk
Zbigniew Stojek

Analiza statystyczna w laboratorium badawczym



4.1. Regresja liniowa zwykła

Jednym z zasadniczych aspektów pracy badawczej jest poszukiwanie modelu matematycznego opisującego relacje między różnymi wielkościami (zmiennymi) opisującymi badane zjawisko lub fragment materii. Dążymy jednocześnie do jak najprostszego matematycznego odwzorowania mechanizmów rządzących badanym zjawiskiem. Uproszczenia dotyczą zarówno formuły (funkcji) matematycznej, którą staramy się odnaleźć w analizowanych zależnościach, jak i sposobu wyznaczania jej szczegółowej postaci. Najprostszą zależnością matematyczną, do której się dąży przy opisie współzależności między mierzonymi wielkościami, jest funkcja liniowa. Dalsze rozważania w tym rozdziale będą zatem w głównej mierze dotyczyć zależności liniowych.

Na istnienie zależności między dwiema zmiennymi może wskazywać wynik analizy wariancji (ANOVA) wraz z testem post-hoc. Stwierdzenie istotnych różnic między wszystkimi możliwymi kombinacjami wartości średnich mierzonej cechy jest bardzo charakterystycznym rezultatem w analizie wariancji. Taki wynik wskazuje na możliwość występowania zależności między zmienną reprezentowaną przez wartości mierzonej cechy (zmienną objaśnianą, zależną, zwykle oznaczaną symbolem y) a czynnikiem, który możemy rozpatrywać jako zmienną niezależną (objaśniającą, zwykle oznaczaną symbolem x), reprezentowaną przez wybrane wartości. Ilościowym potwierdzeniem istnienia zależności liniowej między dwiema zmiennymi jest **współczynnik korelacji liniowej**. Oznacza się go zwykle literą r i oblicza według następującego wzoru:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.1)$$

gdzie: x_i i y_i są współrzędnymi kolejnych punktów w analizowanym zbiorze, n – liczbą punktów, a \bar{x} i \bar{y} to odpowiednio wartości średnie współrzędnych x oraz y .

dla wszystkich punktów. W laboratorium analitycznym x_i i y_i będą na przykład kolejnymi stężeniami analitu i odpowiadającymi im wartościami sygnału analitycznego (odpowiedzi instrumentu). Zależność między y (zmienna zależna) a x (zmienna niezależna) będzie idealnie liniowa, gdy współczynnik korelacji wyniesie 1 albo -1. W analizie chemicznej oczekujemy korelacji lepszych niż 0,99, ale nie zawsze tak dobra korelacja jest możliwa. Jedną z nich jest zwykle analiza śladowa. Wartość bezwzględna współczynnika korelacji znacznie mniejsza od 1 może oznaczać brak liniowej zależności między korelowanymi zmiennymi. Nie znaczy to jednak, że nie ma między nimi żadnej zależności. Przykładowo, jeżeli zależność między y a x jest opisana funkcją kwadratową ($y = x^2$), współczynnik korelacji liniowej może mieć wartość bliską零.

Zależność liniową przedstawiamy ogólnie w postaci równania

$$y = ax + b \quad (4.2)$$

gdzie: a jest współczynnikiem nachylenia prostej (nachyleniem prostej, *slope of a line*), b – wartością przecięcia prostej z osią y (współczynnikiem przecięcia, *y-intercept*). Przyjęta konwencja oznaczeń współczynników równania linii prostej (a dla współczynnika nachylenia oraz b dla współczynnika przecięcia) nie zawsze jest powielana w literaturze. Zdarza się, że linia prosta jest przedstawiana równaniem $y = a + bx$. Dlatego przy wykorzystywaniu wartości współczynników a i b należy zachować ostrożność.

Parametry a i b wyznacza się **metodą najmniejszych kwadratów** (*least square method*). Metoda ta zakłada między innymi, że zmienna niezależna ma charakter nielosowy i jej wartości są ustaloną liczbami. Z kolei charakter zmiennej zależnej jest losowy i opisywany rozkładem normalnym. Konsekwencją tego założenia jest przyjęcie, że niepewności wartości x_i są równe 0 lub są zaniedbywalne w porównaniu do niepewności odpowiadających im wartości y_i . Błąd wyznaczania współczynników zależności funkcyjnej między zmiennymi y i x jest w takiej sytuacji determinowany wyłącznie błędami pomiarowymi wartości zmiennej zależnej. W przypadku regresji liniowej zwykłej (*unweighted linear regression*) metoda najmniejszych kwadratów opiera się dodatkowo na założeniu o identyczności (jednorodności) odchyleń standardowych (lub szerzej – niepewności standardowych) wszystkich wartości y_i . Oznacza to, że wszystkie punkty na linii prostej są jednakowo cenne. W laboratorium przed przystąpieniem do analizy regresji liniowej powinno się więc zweryfikować powyższe upraszczające założenia.

Wskazane założenia można przedstawić za pomocą następujących relacji:

- I. $\frac{u(y_i)}{y_i} \gg \frac{u(x_i)}{x_i}$ dla każdego punktu $i = 1, \dots, n$

II. $u(y_i) = \text{const}$ dla każdego punktu $i = 1, \dots, n$

lub w przypadku bardzo rozległego zakresu wartości zmiennej y

II. $\frac{u(y_i)}{y_i} = \text{const}$ dla każdego punktu $i = 1, \dots, n$

Równanie linii prostej wyznacza się tak, aby suma kwadratów odległości (wzdłuż osi y) wszystkich punktów pomiarowych od poszukiwanej prostej dawała najmniejszą wartość. Procedura ta generuje następujące wyrażenia na współczynnik nachylenia a oraz współczynnik b przecięcia z osią y :

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.3)$$

$$b = \bar{y} - a\bar{x} \quad (4.4)$$

gdzie:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (4.5)$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad (4.6)$$

n jest liczbą punktów wziętych do obliczeń linii prostej.

▲ Przykład 4.1.

Poszukiwanie metodą najmniejszych kwadratów zależności liniowej między wysokością piku woltamperogramu i stężeniem cynku w przygotowanych próbkach wzorcowych.

x (stężenie cynku w mg/l)	0	2	4	6	8	10	12
y (wysokość piku w μA)	0,11	4,90	9,72	14,45	19,07	22,47	24,20

Rozwiązanie

W praktyce obliczenia regresji liniowej przeprowadza się za pomocą wyspecjalizowanych programów komputerowych lub kalkulatora. „Ręczne” wykonanie obliczeń

pozwala jednak pełniej zrozumieć procedurę obliczeniową oraz wpływ poszczególnych wyrażeń na uzyskane wyniki. Kolejne etapy obliczeń przedstawia tabela poniżej.

Nr	x_i	y_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	0	0,11	-6	36	-13,45	180,90	80,70
2	2	4,90	-4	16	-8,66	75,00	34,64
3	4	9,72	-2	4	-3,84	14,75	7,68
4	6	14,45	0	0	0,89	0,79	0,00
5	8	19,07	2	4	5,51	30,36	11,02
6	10	22,47	4	16	8,91	79,39	35,64
7	12	24,20	6	36	10,64	113,21	63,84
$\sum_{i=1}^7$	42	94,92	0	112	0	494,39	233,52

Po podstawieniu wartości odpowiednich sum do wzorów (4.5), (4.6), a następnie (4.3), (4.4) oraz (4.1) otrzymujemy odpowiednio wartości \bar{x} , \bar{y} , a , b oraz r . Równanie wyznaczonej prostej ma więc następującą postać:

$$y = 2,085x + 1,050; r = 0,992.$$



W związku z istnieniem błędów losowych wyznaczone wartości współczynników nachylenia oraz przecięcia na pewno nie pokrywają się z wartościami prawdziwymi. Szerokości przedziałów, w których znajdują się wartości prawdziwe a i b , determinują ich niepewności standardowe (odchylenia standardowe, odpowiednio s_a oraz s_b). Te zaś można tylko oszacować, stosując formalizm propagacji niepewności określony równaniem (2.14), gdyż parametry a i b nie są bezpośrednio mierzone, lecz jako wielkości złożone są wyznaczane na podstawie pomiarów zmiennych x i y (patrz równania 4.3 i 4.4). Przyjmuje się, że niepewność pomiaru korelowanych zmiennych (w szczególności zmiennej zależnej y) można ocenić na podstawie resztowego odchylenia standardowego (*residual standard deviation*). Parametr ten jest miernikiem dyspersji punktów eksperymentalnych wokół dopasowanej prostej. Opisany jest on następującym wzorem:

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} \quad (4.7)$$

gdzie: \hat{y}_i – wartości obliczone na podstawie wyznaczonego równania prostej $\hat{y}_i = ax_i + b$.

Resztowe odchylenie standardowe jest więc głównym czynnikiem kształtującym wielkość niepewności współczynników a i b . Odpowiednie wyrażenia na niej pewności standardowe a i b wyglądają następująco:

$$u(a) = s_a = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (4.8)$$

$$u(b) = s_b = s_{y/x} \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}} \quad (4.9)$$

Współczynnikom a i b przypisujemy rozkład normalny lub, w przypadku nie-wielkich zbiorów punktów, rozkład t -Studenta. Oczekujemy zatem, że wartości prawdziwe a i b , z zadanym prawdopodobieństwem, znajdują się w przedziałach:

$$a \pm ts_a \quad (4.10)$$

oraz

$$b \pm ts_b \quad (4.11)$$

Dla danych z przykładu 4.1, wielkości przedziałów, w których znajdują się prawdziwe wartości współczynników a i b (z prawdopodobieństwem 95%), wynoszą odpowiednio

$$a \pm ts_a = (2,08 \pm 0,30) \mu\text{A} (\text{mg/l})^{-1}$$

oraz

$$b \pm ts_b = (1,05 \pm 2,15) \mu\text{A}$$

Warto zwrócić uwagę, że parametr t odczytany z tabeli rozkładu t -Studenta dla prawdopodobieństwa 95% wynosi 2,571, gdyż liczba stopni swobody dla 7 punktów na linii prostej wynosi $n - 2$, czyli 5.

4.2. Krzywa kalibracyjna. Wyznaczanie stężenia analitu w badanej próbce

Zależność liniowa między sygnałem analitycznym a stężeniem zwykle pełni funkcję krzywej kalibracyjnej (*calibration curve*). Dla większości metod analitycznych teoretyczna krzywa kalibracyjna może być opisana prostszym równaniem $y = ax$.

Wówczas wartość współczynnika b jest równa 0, gdyż brak analitu (ślepa próba, $x = 0$) powinien być równoznaczny z brakiem sygnału ($y = 0$). W praktyce nigdy nie otrzymujemy jednak wartości b równej 0. Niezerowa wartość współczynnika przecięcia z osią y może sugerować istnienie stałego błędu systematycznego w pomiarach. Aby sprawdzić słuszność tego przypuszczenia, należy zawsze obliczać szerokości przedziałów, w których znajdują się prawdziwe wartości współczynników a i b – równania (4.10) i (4.11). Jeśli obliczony przedział dla b zawiera 0, hipotezę o istnieniu stałego błędu systematycznego w pomiarach można zanegować.

Krzywa kalibracyjna najczęściej służy do wyznaczania stężenia analitu w badanej próbce na podstawie pomiaru sygnału dla tej próbki. Odpowiednie wyrażenie na to stężenie uzyskujemy, przekształcając wyznaczone równanie krzywej kalibracyjnej

$$\bar{x}_0 = \frac{\bar{y}_0 - b}{a} \quad (4.12)$$

gdzie: \bar{x}_0 jest średnim stężeniem analitu w badanej próbce odpowiadającym średniej wartości zmierzonego sygnału analitycznego \bar{y}_0 dla tej próbki.

Równanie (4.12) służy również do znalezienia wyrażenia na niepewność standardową (złożoną) x_0 . Postać jego wskazuje, że niepewność standardową tej zmiennej kształtują cztery zasadnicze składowe: niepewność zmiennej zależnej, tj. zmierzonego sygnału dla badanej próbki \bar{y}_0 , niepewności wyznaczonych współczynników regresji a i b oraz korelacja między tymi współczynnikami. Ogólną postać równania na niepewność standardową x_0 można uzyskać na podstawie równania (2.14)

$$u(\bar{x}_0) = \frac{1}{a} \sqrt{u^2(\bar{y}_0) + u^2(b) + \bar{x}_0^2 u^2(a) + 2\bar{x}_0 u(a) u(b) r_{ab}} \quad (4.13)$$

gdzie: $u(a)$, $u(b)$ oraz $u(\bar{y}_0)$ reprezentują odpowiednio: odchylenie standardowe współczynnika a – równanie (4.8), odchylenie standardowe współczynnika b – równanie (4.9) oraz odchylenie standardowe średniej wartości y_0 , natomiast r_{ab} jest współczynnikiem korelacji między a i b .

Aby uniknąć uwzględniania istniejącej korelacji między współczynnikami a i b , można współczynnik b w równaniu (4.12) zastąpić równaniem (4.4); wówczas x_0 uzyskuje następującą postać:

$$\bar{x}_0 = \frac{\bar{y}_0 - \bar{y}}{a} + \bar{x} \quad (4.14)$$

Warto zaznaczyć, że ten matematyczny zabieg eliminuje uwzględnienie korelacji między zmiennymi, nie powodując zmiany wielkości niepewności wyniku końcowego. Niepewność x_0 kształtać będą teraz: sygnał analityczny dla badanej próbki \bar{y}_0 , wartość średnia sygnałów analitycznych dla punktów krzywej kalibracyjnej \bar{y} oraz współczynnik nachylenia krzywej kalibracyjnej a . Nie uwzględnia się w tym zestawie wkładu od niepewności \bar{x} , gdyż zgodnie z założeniami regresji liniowej zwykłej

niepewność stężeń wziętych do konstrukcji krzywej kalibracyjnej x , jest znacznie mniejsza od niepewności pomiaru odpowiadających im sygnałów analitycznych.

Brak korelacji wśród zmiennych w wyrażeniu definiującym x_0 pozwala wykorzystać prostszy formalizm matematyczny propagacji niepewności reprezentowanego równaniem (2.15). Ogólne wyrażenie na niepewność złożoną x_0 przyjmuje teraz następującą postać:

$$u(\bar{x}_0) = \frac{1}{a} \sqrt{u^2(\bar{y}_0) + u^2(\bar{y}) + \frac{(\bar{y}_0 - \bar{y})^2}{a^2} u^2(a)} \quad (4.15)$$

W przypadku regresji liniowej zwykłej poszczególne wyrazy reprezentujące niepewności standardowe w równaniach (4.13) i (4.15) można wyrazić za pomocą resztowego odchylenia standardowego $s_{y/x}$. Wynika to z założenia identyczności odchyleń standardowych punktów wziętych do konstrukcji krzywej kalibracyjnej i pozwala uzyskać uniwersalną postać wyrażenia na niepewność złożoną wielkości x_0 , niezależnie od sposobu przedstawienia x_0 .

$$u(\bar{x}_0) = s_{x_0} = \frac{s_{y/x}}{a} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_0 - \bar{y})^2}{a^2 \sum_{i=1}^n (x_i - \bar{x})^2}} \quad (4.16)$$

W powyższym wzorze m jest liczbą pomiarów sygnału analitycznego dla bieżącej próbki, n liczbą punktów krzywej kalibracyjnej, a $s_{y/x}$ jest resztowym odchyleniem standardowym danym wzorem (4.7).

Wielkość x_0 podlega rozkładowi t -Studenta, jeżeli wszystkie jej źródła niepewności mają charakter losowy. Niepewność standardową x_0 możemy utożsamiać wtedy z odchyleniem standardowym tego rozkładu. Odpowiedni przedział, w którym z założonym prawdopodobieństwem znajduje się wartość prawdziwa x_0 , można więc zapisać jako:

$$x_0 \pm ts_{x_0} \quad (4.17)$$

gdzie: s_{x_0} jest odchyleniem standardowym x_0 , a wartość parametru t odczytujemy dla $n - 2$ stopni swobody. Warto zauważyć, że szerokość przedziału ufności (2 · s_{x_0}) będzie stawała się coraz większa przy wzroście zarówno liczby pomiarów m , jak i liczby punktów n wziętych do konstrukcji krzywej kalibracyjnej.

▲ Przykład 4.2.

Przeprowadzono woltamperometryczną analizę dwóch próbek z wykorzystaniem równania regresji liniowej obliczonego w przykładzie 4.1 ($y = 2,085x + 1,050$, $r = 0,992$). Uzyskano następujące wartości sygnału prądowego:

- 1) y_0 : 4,50; 4,63; 4,54 μA $\bar{y}_{0;1} = 4,56 \mu\text{A}$
 2) y_0 : 23,41; 24,20; 22,59 μA $\bar{y}_{0;2} = 23,40 \mu\text{A}$

Wyznaczyć wartości stężeń analitu w badanych próbkach wraz z przedziałami ufności.

Rozwiążanie

Wartości stężeń analitu w badanych próbkach obliczamy na podstawie wzoru (4.12)

$$\bar{x}_{0;1} = \frac{\bar{y}_{0;1} - b}{a} = \frac{4,56 - 1,050}{2,085} = 1,68 \text{ mg/l}$$

$$\bar{x}_{0;2} = \frac{\bar{y}_{0;2} - b}{a} = \frac{23,40 - 1,050}{2,085} = 10,72 \text{ mg/l}$$

Odchylenia standardowe wyznaczonych stężeń obliczamy na podstawie wzoru (4.16) oraz danych z tabeli w przykładzie 4.1

$$s_{x_{0;1}} = \frac{s_{y/x}}{a} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_{0;1} - \bar{y})^2}{a^2 \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{1,23}{2,085} \sqrt{\frac{1}{3} + \frac{1}{7} + \frac{(4,56 - 13,56)^2}{(2,085)^2 \cdot 112}} = 0,47 \text{ mg/l}$$

$$s_{x_{0;2}} = \frac{s_{y/x}}{a} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_{0;2} - \bar{y})^2}{a^2 \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{1,23}{2,085} \sqrt{\frac{1}{3} + \frac{1}{7} + \frac{(23,40 - 13,56)^2}{(2,085)^2 \cdot 112}} = 0,48 \text{ mg/l}$$

Dla poziomu prawdopodobieństwa 95% przedziały, w których znajdują się wartości prawdziwe stężeń analitu w badanych próbkach, są następujące:

$$x_{0;1} \pm ts_{x_{0;1}} = (1,7 \pm 1,2) \text{ mg/l}$$

$$x_{0;2} \pm ts_{x_{0;2}} = (10,7 \pm 1,2) \text{ mg/l}$$

Parametr t odczytany z tabeli rozkładu t -Studenta dla prawdopodobieństwa 95% i $n - 2$, czyli 5 stopni swobody wynosi 2,571.

Warto zauważyć, że szerokości przedziałów są bardzo zbliżone (a po zaokrągleniu do dwóch cyfr znaczących identyczne). Jest to konsekwencja założenia, że odchylenia standardowe wszystkich punktów pomiarowych krzywej kalibracyjnej są identyczne. Sytuacja się zmieni, jeśli się okaże, że punktom wziętym do konstrukcji krzywej kalibracyjnej odpowiadają różne odchylenia standardowe, czyli że w sensie statystycznym mają one różne wagи.



instrumentalnych, że wyznaczenie odchylenia standardowego tła nie jest możliwe lub nie jest uzasadnione. Przykładowo, w woltamperometrii ze wstępny założeniem analitu tło jest często duże. Poziom jego zmienia się znacznie w kolejnych krzywych, ale zawsze ma w miarę płaski przebieg względem potencjału. Na takim płaskim tle sygnał analitu jest często wyraźnie ukształtowany i relatywnie łatwy mierzalny. W tej sytuacji użycie odchylenia standardowego tła jest dyskusyjne i bardziej wiarygodne wartości granicy wykrywalności i granicy oznaczalności otrzyma się, wykorzystując krzywą dodatku wzorca. Sześciokrotna wartość odchylenia standardowego punktu przecięcia krzywej z osią stężeń s_{x_E} określa granicę wykrywalności, a dziesięciokrotna wartość tego odchylenia standardowego odpowiada granice oznaczalności. Dla danych z przykładu 4.3 granice wykrywalności oraz oznaczalności wynoszą odpowiednio: $6s_{x_E} = 1,6$ oraz $10s_{x_E} = 2,6 \mu\text{g/l}$ (wartość s_{x_E} obliczona na podstawie równania (4.19) wynosi $0,26 \mu\text{g/l}$).

Metoda dodatku wzorca, obok oczywistych zalet związanych z wyeliminowaniem efektów matrycy czy możliwości zastosowania do określania granicy wykrywalności, posiada także pewne wady. Do najważniejszych należy zaliczyć: trudność w automatyzacji procedury, większe zużycie analizowanej próbki, mniejszą (z opartą na ekstrapolacji) dokładność oznaczania stężenia analitu w badanej próbce i, wreszcie, większą niepewność wynikającą z założenia, że sygnał mierzony bez dodatku wzorca wynika tylko z obecności analitu.

4.4. Regresja liniowa ważona (Y)

Przyjrzyjmy się teraz dokładnie wszystkim wynikom wziętym do konstrukcji krzywej kalibracyjnej z przykładu 4.1 zebranym w tabeli 4.1.

Tabela 4.1. Pełna lista danych do konstrukcji krzywej kalibracyjnej z przykładu 4.1 wraz z charakterystyką statystyczną

Nr	$x_i [\text{mg/l}]$	$y_i [\mu\text{A}]$	$\bar{y}_i [\mu\text{A}]$	$s(\bar{y}_i) [\mu\text{A}]$	w_i
1	0	0,09; 0,11; 0,13	0,11	0,012	6,3
2	2	4,90; 4,98; 4,81	4,90	0,049	0,35
3	4	9,72; 9,60; 9,84	9,72	0,069	0,18
4	6	14,35; 14,40; 14,60	14,45	0,076	0,14
5	8	19,11; 19,40; 18,70	19,07	0,20	0,020
6	10	22,51; 21,89; 23,00	22,47	0,32	0,0081
7	12	24,22; 25,00; 23,38	24,20	0,47	0,0038

W piątej kolumnie tabeli 4.1 są umieszczone odchylenia standardowe odpowiadające kolejnym seriom pomiarowym sygnału analitycznego. Łatwo zauważyć, że odchylenia standardowe rosną wraz ze wzrostem stężenia. Punkty dla większych wartości x powinny mieć zatem mniejszy wpływ na nachylenie i przecięcie wyznaczonej linii prostej. Ten różny wpływ punktów może być uwzględniony poprzez przyporządkowanie kolejnym wartościami \bar{y}_i odpowiednich wag. Ilościowo wagę i -tego punktu (w_i) można określić odwrotnością kwadratu odchylenia standardowego (niepewności standardowej) wartości zmiennej y tego punktu, czyli $w_i = \frac{1}{u^2(\bar{y}_i)} = \frac{1}{s^2(\bar{y}_i)}$.

Wartości wag umieszczone w ostatniej kolumnie tabeli 4.1 zostały obliczone na podstawie następującego wzoru:

$$w_i = \frac{n [s(\bar{y}_i)]^{-2}}{\sum_{i=1}^n [s(\bar{y}_i)]^{-2}} \quad (4.21)$$

Powyższy wzór definiuje znormalizowane wagi i jest najczęściej stosowaną definicją wag punktów wziętych do obliczeń regresyjnych w chemii analitycznej. Analiza tej formuły pozwala wyprowadzić przydatne kryterium poprawności obliczeń wag, tj. $\sum_{i=1}^n w_i = n$. Wprowadzenie wag do metody najmniejszych kwadratów prowadzi do uzyskania wyrażeń będących uogólnionymi wersjami wzorów zwykłej regresji liniowej. Ten wariant regresji liniowej nazywamy regresją liniową ważoną (Y) (*Y weighted linear regression*). Współczynniki zależności liniowej są wyznaczane tak, aby suma kwadratów ważonych odległości punktów od linii prostej była jak najmniejsza. Odpowiednie wyrażenia na współczynniki a i b oraz ich niepewności standardowe (odchylenia standardowe) wyznaczone metodą regresji ważonej (Y) mają następujące postaci:

$$a = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)(\bar{y}_i - \bar{y}_w)}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2} \quad (4.22)$$

$$b = \bar{y}_w - a\bar{x}_w \quad (4.23)$$

$$s_a = s_a = \frac{s_{y/x,w}}{\sqrt{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}} \quad (4.24)$$

$$u(b) = s_b = s_{y/x,w} \sqrt{\frac{\sum_{i=1}^n w_i x_i^2}{n \sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}}$$

gdzie: x_i oraz \bar{y}_i oznaczają odpowiednio wartość zmiennej niezależnej x oraz
wartość średnią zmiennej y dla i -tego punktu, a \bar{x}_w i \bar{y}_w oraz $s_{y/x,w}$ dane są wzorami

$$\bar{x}_w = \sum_{i=1}^n \frac{w_i x_i}{n}$$

$$\bar{y}_w = \sum_{i=1}^n \frac{w_i \bar{y}_i}{n}$$

$$s_{y/x,w} = \sqrt{\frac{\sum_{i=1}^n w_i (y_i - ax_i - b)^2}{n-2}}$$

Parametr $s_{y/x,w}$ jest resztowym odchyleniem standardowym w regresji ważonej

▲ Przykład 4.4.

Obliczyć parametry zależności liniowej metodą regresji ważonej (Y) dla punktów pomiarowych umieszczonych w tabeli 4.1.

Kolejne etapy obliczeń zestawiono w tabeli poniżej.

Nr	$w_i x_i$	$w_i \bar{y}_i$	$(x_i - \bar{x}_w)$	$w_i (x_i - \bar{x}_w)^2$	$(\bar{y}_i - \bar{y}_w)$	$w_i (x_i - \bar{x}_w)(\bar{y}_i - \bar{y}_w)$
1	0,000	0,693	-0,364	0,834	-0,867	1,986
2	0,700	1,715	1,636	0,937	3,923	2,247
3	0,720	1,750	3,636	2,380	8,743	5,723
4	0,840	2,023	5,636	4,447	13,473	10,631
5	0,160	0,381	7,636	1,166	18,093	2,763
6	0,081	0,182	9,636	0,752	21,493	1,678
7	0,046	0,092	11,636	0,515	23,223	1,027
$\sum_{i=1}^7$	2,547	6,836	39,453	11,031	88,084	26,055

Po podstawieniu wartości sum $w_i x_i$ oraz $w_i \bar{y}_i$ do równań (4.26) i (4.27) otrzymujemy wartości średnich ważonych \bar{x}_w oraz \bar{y}_w ,

$$\bar{x}_w = \frac{2,547}{7} = 0,364 \text{ mg/l}$$

$$\bar{y}_w = \frac{6,836}{7} = 0,977 \mu\text{A}$$

Wartości współczynników a i b oraz wartości ich niepewności standardowych wyznaczamy na podstawie wzorów (4.22) i (4.23) oraz (4.24) i (4.25). Wynoszą one odpowiednio:

$$a = \frac{26,055}{11,031} = 2,362 \mu\text{A} (\text{mg/l})^{-1}$$

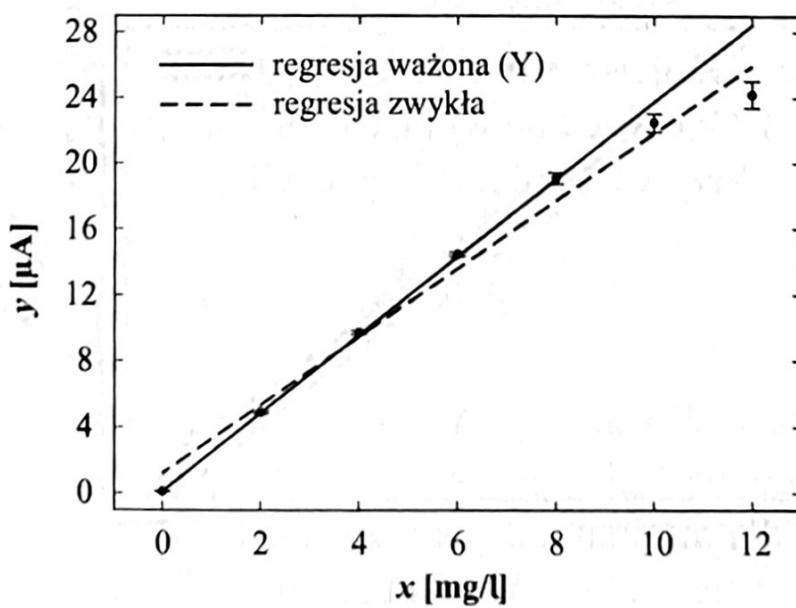
$$b = 0,977 - 2,362 \cdot 0,364 = 0,117 \mu\text{A}$$

$$u(a) = s_a = \frac{0,136}{\sqrt{11,031}} = 0,041 \mu\text{A} (\text{mg/l})^{-1}$$

$$u(b) = s_b = 0,136 \sqrt{\frac{11,957}{7 \cdot 11,031}} = 0,054 \mu\text{A}$$

Wyznaczenie wartości $u(a)$ oraz $u(b)$ wymagało obliczenia wcześniej wartości resztowego odchylenia standardowego ($s_{y/x,w} = 0,136 \mu\text{A}$) oraz sumy: $\sum_{i=1}^7 w_i x_i^2 = 11,957$.

Otrzymane wartości współczynników a i b nie różnią się znacznie od wartości uzyskanych na podstawie zwykłej regresji. Obie proste o równaniach $y = 2,085x + 1,050$ (metoda regresji liniowej zwykłej, przykład 4.1) oraz $y = 2,362x + 0,117$ (metoda regresji liniowej ważonej) przedstawione są razem na rysunku 4.2.



Rys. 4.2. Porównanie prostych wyznaczonych metodą regresji liniowej zwykłej (linia przerywana) i ważonej Y (linia ciągła)

Na rysunku 4.2 widzimy wyraźnie różnice między dwoma wariantami metody regresji liniowej. W przypadku metody regresji liniowej zwykłej wszystkie punkty eksperymentalne w takim samym stopniu decydują o przebiegu prostej (tj. o wartościach jej współczynników nachylenia i przecięcia). Fakt ten jest bezpośrednią konsekwencją założenia statystycznej identyczności odchyлеń standardowych wartości zmiennej y (sygnału analitycznego). Wyznaczona prosta jest zatem optymalnie dopasowana do wszystkich punktów eksperymentalnych. W przeciwnieństwie do metody regresji liniowej zwykłej metoda ważona (Y) różnicuje wpływ poszczególnych punktów na przebieg wyznaczanej prostej. Innymi słowy, w metodzie regresji liniowej ważonej (Y) niejednorodność niepewności standardowych przypisanych wartości zmiennej y powoduje zróżnicowany wpływ kolejnych punktów na parametry dopasowanej prostej. Ilościową miarą wpływu danego punktu na przebieg prostej jest przypisana mu waga. Punkty, a ścisłeji wartości zmiennej y , wyznaczone z małą precyją, tj. charakteryzujące się dużymi odchyleniami standardowymi i w konsekwencji małymi wagami, mają mniejszy wpływ na wartości współczynników regresji niż punkty, które zostały wyznaczone z większą precyją, a ścisłeji charakteryzują się większymi wagami.

Dla danych rozważanych w tym przykładzie (tabela 4.1) założenie o identyczności odchyłeń standardowych zmiennej y jest ewidentnie niespełnione. Odchylenia standardowe skrajnych punktów różnią się nawet o rząd wielkości. Metoda regresji liniowej ważonej (Y) jest zatem w tym przypadku uzasadnionym wyborem schematu obliczeniowego. Łatwo zauważyc na rysunku 4.2, że prosta regresji ważonej jest w znikomym stopniu dopasowana do trzech ostatnich punktów eksperymentalnych. Punktom tym przypisane są bowiem bardzo niewielkie wagi. O przebiegu prostej decydują więc pozostałe, precyzyjniej wyznaczone punkty.



Istnienie różnych wag może mieć duży wpływ na zastosowania krzywej kalibracyjnej. Jednym z ważniejszych zastosowań jest wyznaczanie stężenia analitycznego w badanej próbce (x_0). Odpowiednie wyrażenie uzyskujemy po przekształceniu wyznaczonego równania krzywej kalibracyjnej, czyli

$$\bar{x}_0 = \frac{\bar{y}_0 - b}{a} \quad (4.2)$$

W przypadku regresji ważonej (Y) konieczne jest uzyskanie serii pomiarów sygnału analitycznego dla badanej próbki. Wartość \bar{y}_0 musi być więc średnią arytmetyczną obliczoną dla minimum 3 pomiarów.

Do wyznaczenia przedziału, w którym znajduje się wartość prawdziwa stężenia odczytanego z krzywej kalibracyjnej, niezbędne jest oszacowanie odchylenia standardowego punktu x_0 . Postępując analogicznie jak w przypadku regresji zwykłej

uzyskujemy następujące uogólnione wyrażenie na odchylenie standardowe x_0 (niepewność złożoną x_0):

$$u(\bar{x}_0) = s_{x_0,w} = \frac{s_{y/x,w}}{a} \sqrt{\frac{1}{w_0} + \frac{1}{n} + \frac{(\bar{y}_0 - \bar{y}_w)^2}{a^2 \sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}} \quad (4.30)$$

gdzie: $s_{y/x,w}$ jest resztowym odchyleniem standardowym w regresji ważonej – równanie (4.28), a w_0 jest wagą dla pomiarów sygnału analitycznego badanej próbki zdefiniowaną następującym wzorem:

$$w_0 = \frac{n [s(\bar{y}_0)]^{-2}}{\sum_{i=1}^n [s(\bar{y}_i)]^{-2}} \quad (4.31)$$

Spójrzmy teraz, jakie otrzymamy przedziały dla prawdziwego stężenia analitu w badanych próbkach, które rozważaliśmy w przykładzie 4.2.

Dla \bar{y}_0 równe 4,56 i 23,40 μA (średnie z trzech pomiarów) szerokości przedziałów dla prawdziwego stężenia analitu są teraz następujące:

$(1,88 \pm 0,21) \text{ mg/l}$ i $(9,9 \pm 2,4) \text{ mg/l}$ (w porównaniu do wyników $(1,7 \pm 1,2) \text{ mg/l}$ i $(10,7 \pm 1,2) \text{ mg/l}$ otrzymanych dla regresji zwykłej).

O ile wartości x_0 obliczone na podstawie regresji zwykłej i ważonej (Y) są w miarę zbliżone, to szerokości przedziałów ufności różnią się znacznie. Przy wykorzystaniu regresji ważonej przedział dla prawdziwego stężenia analitu jest znacznie węższy dla próbki o małym stężeniu i znacznie szerszy dla drugiej próbki. Jest to rezultat znacznie większych wag dla małych stężeń krzywej kalibracyjnej w porównaniu do dużych stężeń. W konsekwencji niepewność oznaczeń w obszarze większych stężeń jest znaczco większa.

5. Regresja liniowa ważona (X,Y) z uwzględnieniem niepewności obu zmiennych

Analizujmy jeszcze raz wyniki wzięte do konstrukcji krzywej kalibracyjnej w przykładzie 4.1, przedstawione wcześniej w tabeli 4.1, ale uzupełnione teraz o wartości niepewności standardowych stężeń analitu w przygotowanych próbkach wzorowych. Pełny zestaw danych eksperymentalnych prezentuje tabela 4.2. W zestawie pominięto pomiary dla ślepej próby ($x = 0$), gdyż jej natura (brak analitu) umożliwia wyznaczenie niepewności standardowej stężenia analitu.

Tabela 4.2. Dane do konstrukcji krzywej kalibracyjnej z przykładu 4.1 uzupełnione o wartości niepewności standardowych stężeń analitu oraz wartości względnych niepewności standardowych wartości zmiennych x i y

Nr	x_i [mg/l]	$u(x_i)$ [mg/l]	$u(x_i) / x_i$	\bar{y}_i [μA]	$s(\bar{y}_i)$ [μA]	$s(\bar{y}_i) / \bar{y}_i$
1	2,000	0,022	0,011	4,90	0,049	0,010
2	4,000	0,044	0,011	9,72	0,069	0,0071
3	6,000	0,066	0,011	14,45	0,076	0,0053
4	8,000	0,088	0,011	19,07	0,20	0,010
5	10,00	0,11	0,011	22,47	0,32	0,014
6	12,00	0,13	0,011	24,20	0,47	0,019

Tabela 4.2 zawiera kompletny obraz danych eksperymentalnych do konstrukcji krzywej kalibracyjnej, czyli do poszukiwania liniowej zależności między najbardziej reprezentatywnymi wartościami pomiaru sygnału analitycznego i odpowiadającymi im wartościami stężeń analitu w przygotowanych próbkach wzorcowych. Każda z wartości korelowanych zmiennych jest teraz dodatkowo scharakteryzowana niepewnością standardową. W przypadku zmiennej zależnej niepewność standardowa została oszacowana metodą A, tj. na podstawie wielkości rozproszenia losowego wyników pomiaru (por. tabela 2.1). W przypadku zmiennej niezależnej oszacowane niepewności standardowe stanowią złożenie kilku niezależnych źródeł niepewności wynikających z przyjętego schematu przygotowania próbek. Procedura przygotowania i -tego roztworu wzorcowego polegała w tym przypadku na odpowiednim rozcieńczeniu wzorca podstawowego zgodnie z równaniem: $x_i = C \cdot V_p / V_i$. W takim schemacie postępowania można wyróżnić trzy źródła niepewności: stężenie analitu C w podstawowym roztworze wzorcowym (dane np. z certyfikatu), pobranie objętości V_p próbki wzorca podstawowego (błąd graniczny np. pipety), odmierzanie całkowitej objętości V_k roztworu wzorcowego (błąd graniczny np. kolby miarowej). Postać iloczynowa równania do wyznaczania wartości zmiennej x oraz charakterystyka błędów poszczególnych składowych w tym wzorze prowadzi do wniosku, że każde z tych źródeł ma taki sam wkład względny do złożonej niepewności stężenia. Prawidłowość ta tłumaczy stałą wartość względnej niepewności złożonej kolejnych stężeń analitu i w konsekwencji proporcjonalny wzrost wartości niepewności dla coraz bardziej stężonych roztworów wzorcowych. Tabela 4.2 jest również wzbogacona o kolumnę zawierającą wartości względnej niepewności standardowej kolejnych wartości zmiennej y . Zestawienie tych wartości z odpowiadającymi im wartościami względnych niepewności zmiennej x pozwala teraz dodatkowo zweryfikować założenie o zaniedbywanym charakterze wielkości niepewności wyznaczania wartości x_i w odniesieniu do niepewności przypisanych wartościom

zmiennej y . Na etapie weryfikacji tego założenia wykorzystuje się niepewności względne, aby porównywany efekt zmienności w wartościach mierzonych/wyznaczanych uniezależnić od charakteru (istoty fizycznej) zmiennej. Założenie jest spełnione, jeśli wartości względnych niepewności zmiennej y są co najmniej 10 razy większe od odpowiadających im wartości względnych niepewności zmiennej x . W rozważanym przykładzie nie jest to jednak spełnione i wymagane jest włączenie tego efektu (tj. niepewności skali x) do obliczeń założonej zależności liniowej $y = ax + b$, czyli uzyskania najbardziej prawdopodobnych ocen wartości współczynników a i b oraz ich niepewności standardowych $u(a)$ i $u(b)$.

Jeżeli oba upraszczające założenia w całym zakresie wartości zmiennych x i y nie są spełnione, wymagane jest zastosowanie do obliczeń ogólnego schematu regresji liniowej ważonej (X,Y) uwzględniającej niepewności obydwu skal oraz istotną niejednorodność niepewności mierzonych wartości zmiennej zależnej. Do obliczeń współczynników poszukiwanej funkcji zostaną wykorzystane uśrednione wartości współrzędnych kolejnych punktów eksperymentalnych (x_i, \bar{y}_i) oraz przypisane im wagi W różnicujące wpływ poszczególnych punktów na przebieg poszukiwanej zależności i będące funkcją niepewności zarówno wartości zmiennej zależnej, jak i niezależnej.

Ilościowo wagę ogólną i -tego punktu W_i uwzględniającą niepewności zarówno wartości zmiennej y , jak i x , określać może odwrotność kwadratu niepewności wynikającej z błędu dopasowania punktów eksperymentalnych do założonej funkcji, czyli tzw. błędu reszty regresyjnej. Dla i -tego punktu błąd resztowy wyrazić można jako

$$e_i = \bar{y}_i - ax_i - b \quad (4.32)$$

A zatem,

$$W_i = \frac{1}{u^2(e_i)} = \frac{1}{u^2(\bar{y}_i) + a^2 u^2(x_i)} = \frac{w(x_i)w(\bar{y}_i)}{w(x_i) + a^2 w(\bar{y}_i)} \quad (4.33)$$

gdzie:

$$w(x_i) = \frac{1}{u^2(x_i)} \quad (4.34)$$

$$w(\bar{y}_i) = \frac{1}{u^2(\bar{y}_i)} \quad (4.35)$$

Wyrażenie na W_i zakłada brak korelacji między niepewnościami standardowymi zmiennej x i y . Analiza tej formuły uwidacznia ponadto albo kooperatywny, albo przeciwny charakter udziału wag współrzędnych punktu w wartości W_i . Jeżeli niepewności zmiennej x i y są istotne, ale kierunki ich zmian są przeciwnie (duża niepewność wartości y , przy znikomej niepewności odpowiadającej jej wartości x_i i odwrotnie), efekt zróżnicowania punktów pod wpływem niepewności

obydwu zmiennych zostaje wygaszony. Punkty stają się równocenne i rezultaty obliczeń jakościowo zbliżają się do wyników uzyskanych metodą regresji liniowej zwykłej.

Wprowadzenie ogólnych wag do metody najmniejszych kwadratów prowadzi do sformułowania najbardziej ogólnego formalizmu metody regresji liniowej, który nazywamy regresją liniową ważoną (X,Y) (X, Y weighted linear regression). W przypadku granicznym, gdy niepewności zmiennej x można uznać za nieistotne, wówczas ogólna waga W_i liczbowo dąży do wartości $w(\bar{y}_i)$, a metoda regresji liniowej ważonej (X,Y) jest realizowana według bardziej uproszczonego schematu regresji liniowej ważonej (Y). Jeśli dodatkowo jeszcze wartości niepewności zmiennej y można uznać za jednorodne, wówczas wartości W_i są stałe (umownie równe 1), a formalizm obliczeń regresyjnych sprowadza się do najbardziej uproszczonego schematu regresji liniowej zwykłej.

Odpowiednie wyrażenia na współczynniki a i b oraz ich niepewności standarde (odchylenia standardowe) wyznaczone metodą regresji ważonej (X,Y) mają następujące postaci [22]:

$$a = \frac{\sum_{i=1}^n W_i \lambda_i (\bar{y}_i - \bar{y}_w)}{\sum_{i=1}^n W_i \lambda_i (x_i - \bar{x}_w)} \quad (4.36)$$

$$b = \bar{y}_w - a \bar{x}_w \quad (4.37)$$

$$u(a) = \frac{s_{y/x,W}}{\sqrt{\sum_{i=1}^n W_i (X_i - \bar{X})^2}} \quad (4.38)$$

$$u(b) = s_{y/x,W} \sqrt{\frac{1}{\sum_{i=1}^n W_i} + \frac{\bar{X}^2}{\sum_{i=1}^n W_i (X_i - \bar{X})^2}} \quad (4.39)$$

gdzie:

$$\lambda_i = W_i \left[\frac{(x_i - \bar{x}_w)}{w(\bar{y}_i)} + \frac{a(\bar{y}_i - \bar{y}_w)}{w(x_i)} \right] \quad (4.40)$$

$$\bar{x}_w = \frac{\sum_{i=1}^n W_i x_i}{\sum_{i=1}^n W_i} \quad (\text{wartość średniej ważonej dla eksperymentalnych wartości } x) \quad (4.41)$$

$$\bar{y}_w = \frac{\sum_{i=1}^n W_i \bar{y}_i}{\sum_{i=1}^n W_i} \text{ (wartość średniej ważonej dla eksperymentalnych wartości } y) \quad (4.42)$$

$$\bar{X} = \frac{\sum_{i=1}^n W_i X_i}{\sum_{i=1}^n W_i} \text{ (wartość średnia oczekiwanych wartości } x) \quad (4.43)$$

$$\bar{Y} = \frac{\sum_{i=1}^n W_i \bar{Y}_i}{\sum_{i=1}^n W_i} \text{ (wartość średnia oczekiwanych wartości } y) \quad (4.44)$$

$$X_i = \bar{x}_w + \lambda_i \text{ (wartość oczekiwana } x_i \text{, wartość dopasowana metodą najmniejszych kwadratów (*least-squares-adjusted value*))} \quad (4.45)$$

$$\bar{Y}_i = \bar{y}_w + a\lambda_i \text{ (wartość oczekiwana } \bar{y}_i \text{, wartość dopasowana metodą najmniejszych kwadratów (*least-squares-adjusted value*))} \quad (4.46)$$

$$s_{y/x,w} = \sqrt{\frac{\sum_{i=1}^n W_i (\bar{y}_i - ax_i - b)^2}{n-2}} \text{ (resztowe odchylenie standardowe)} \quad (4.47)$$

Uważna analiza wyrażenia na współczynnik nachylenia wskazuje na jego niewiąną postać, czyli brak możliwości wyznaczenia jego wartości w jednym kroku. Proces poszukiwania najbardziej prawdopodobnej wartości tego współczynnika przebiega iteracyjnie. Schemat postępowania można ująć w 7 krokach:

- 1) wybór przyblizonej wartości współczynnika nachylenia (np. dobrym początkowym przybliżeniem a będzie wartość uzyskana metodą regresji liniowej zwykłej);
- 2) wyznaczenie wartości wag $w(x_i)$ oraz $w(\bar{y}_i)$ dla każdego punktu na podstawie wartości niepewności standardowych odpowiednio x_i i \bar{y}_i – równanie (4.34) i (4.35);
- 3) wyznaczanie wartości wag ogólnych W_i dla każdego punktu na podstawie wartości uzyskanych w krokach 1 i 2 – równanie (4.33);
- 4) wyznaczanie wartości średnich ważonych dla eksperymentalnych wartości zmiennej x i y , tj. \bar{x}_w i \bar{y}_w na podstawie równań (4.41) i (4.42), za pomocą zgromadzonych danych eksperymentalnych (x_i, \bar{y}_i) ;

- 5) obliczenie wartości współczynników λ_i dla każdego punktu – równanie (4.40);
- 6) wyznaczanie lepszego przybliżenia wartości współczynnika a na podstawie wyznaczonych wartości W_i , \bar{x}_W i \bar{y}_W oraz λ_i – równanie (4.36);
- 7) powtarzanie obliczeń w krokach 3–6 z nowouzyskaną wartością a aż do pojawienia się satysfakcjonującej zbieżności w wartościach a , tj. do momentu, kiedy różnice między wartościami a kolejnych iteracji nie będą przekraczać pewnej ustalonej tolerancji np. 10^{-15} .

Po uzyskaniu zbieżności w wyznaczanych wartościach współczynnika a , kolejne kroki obejmują:

- 1) wyznaczenie współczynnika przecięcia b na podstawie najbardziej reprezentatywnej wartości a oraz wartości \bar{x}_W i \bar{y}_W – równanie (4.37);
- 2) wyznaczenie wartości oczekiwanych X_i , Y_i na podstawie równań (4.45) i (4.46);
- 3) wyznaczanie oczekiwanych wartości średnich \bar{X} , \bar{Y} za pomocą danych uzyskanych w kroku 2 oraz wartości wag W_i – równania (4.43) i (4.44);
- 4) szacowanie niepewności współczynnika nachylenia a – z równania (4.38), a następnie współczynnika przecięcia b – z równania (4.39), za pomocą danych uzyskanych w krokach 2 i 3 oraz wartości resztowego odchylenia standardowego $s_{y/x,W}$ wyznaczonej z równania (4.47).

▲ Przykład 4.5.

Wyznaczyć parametry zależności liniowej metodą regresji ważonej (X, Y) dla punktów pomiarowych umieszczonych w tabeli 4.2.

Przed przystąpieniem do obliczeń regresyjnych wyznaczamy wagi dla kolejnych wartości zmiennych x i y oraz przybliżone wartości wag ogólnych W_i na podstawie przybliżonej wartości współczynnika nachylenia uzyskanej metodą regresji liniowej zwykłej. Wstępne dane są przedstawione poniżej w formie tabelarycznej

Nr	$w(x_i)$	$w(\bar{y}_i)$	W_i
1	2066,12	416,49	231,50
2	516,53	210,04	80,42
3	229,57	173,13	43,40
4	129,13	25,00	14,14
5	82,64	9,77	6,65
6	59,17	4,53	3,48

Do obliczeń wykorzystano przybliżoną wartość współczynnika nachylenia, $1,991 \mu\text{A} (\text{mg/l})^{-1}$. Następnie iteracyjnie, zgodnie z opisaną wcześniej procedurą

wyznaczono zestaw najbardziej reprezentatywnych wartości parametrów poszukiwanej zależności liniowej. Wynoszą one odpowiednio:

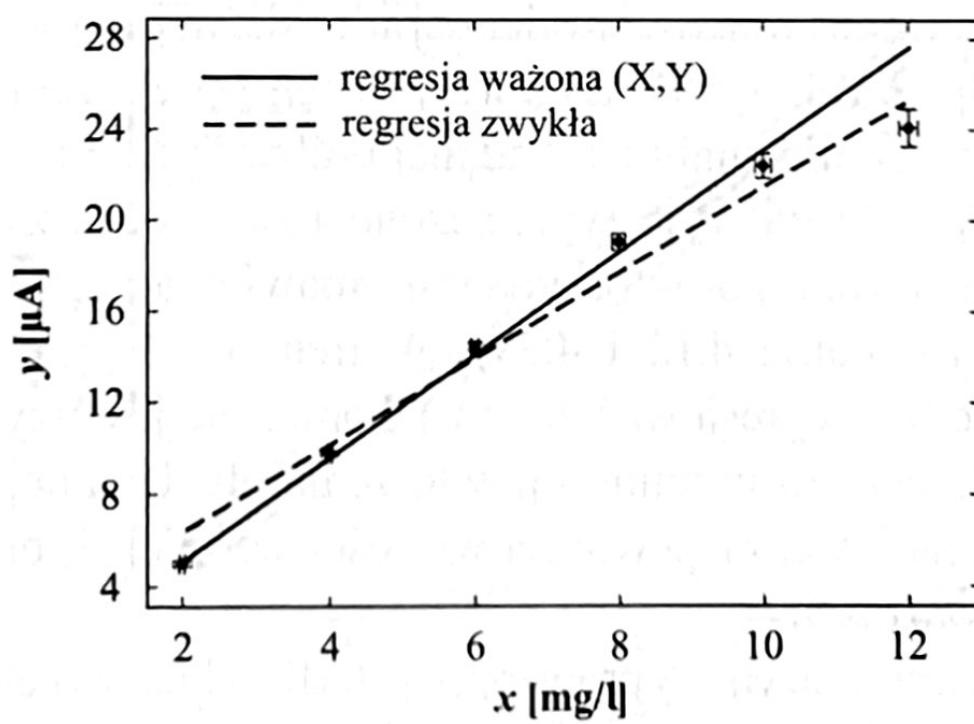
$$a = 2,256 \mu\text{A} (\text{mg/l})^{-1}$$

$$b = 0,492 \mu\text{A}$$

$$u(a) = 0,096 \mu\text{A} (\text{mg/l})^{-1}$$

$$u(b) = 0,38 \mu\text{A}$$

Otrzymane wartości współczynników a i b wykazują znaczne odstępstwa od wartości uzyskanych na podstawie zwykłej regresji ($a = 1,991 \mu\text{A} (\text{mg/l})^{-1}$, $b = 1,865 \mu\text{A}$). Znaczące różnice można też zaobserwować dla wartości niepewności standardowych wyznaczonych współczynników prostej ($u(a) = 0,096 \mu\text{A} (\text{mg/l})^{-1}$ oraz $u(b) = 0,38 \mu\text{A}$ w metodzie ważonej (X,Y) wobec $u(a) = 0,145 \mu\text{A} (\text{mg/l})^{-1}$ oraz $u(b) = 1,13 \mu\text{A}$ w metodzie regresji zwykłej). W tym przypadku metoda regresji liniowej ważonej (X,Y) generuje znaczaco węższe przedziały niepewności współczynników a i b , ponieważ ich wyznaczanie w niewielkim stopniu uwzględniało ostatnie dwa punkty o najniższych wagach, które dodatkowo zdecydowanie odstają od prostoliniowego przebiegu. Przebieg prostej wyznaczony metodą regresji liniowej zwykłej uwzględnia wszystkie punkty z jednakową wagą, co skutkuje stosunkowo dużym rozrzutem punktów wokół prostej oraz słabą korelacją liniową punktów ($r < 0,99$). Obie proste: $y = 1,991x + 1,865$ (metoda regresji liniowej zwykłej) oraz $y = 2,256x + 0,492$ (metoda regresji liniowej ważonej (X,Y)) przedstawione są razem na rysunku 4.3.



Rys. 4.3. Porównanie prostych wyznaczonych metodą regresji liniowej zwykłej – linia przerywana i ważonej (X,Y) – linia ciągła

Na rysunku 4.3 widać wyraźnie różnice między dwoma skrajnymi wariantami metody regresji liniowej. W metodzie regresji liniowej zwykłej wszystkie punkty

eksperymentalne są równocenne i w takim samym stopniu decydują o przebiegu prostej. Metoda regresji liniowej ważonej (X, Y), podobnie jak metoda ważona (Y), różnicuje wpływ poszczególnych punktów na przebieg wyznaczanej prostej. W przeciwnieństwie jednak do jednowymiarowego ważenia w metodzie ważonej (Y), tym razem każdy punkt jest ważony w dwóch wymiarach. Poszukiwane jest zatem najlepsze dopasowanie zależności liniowej nie do punktów czy odcinkowych zakresów, ale do obszarów możliwych wartości zmiennej x i y , których pola są wytyczone przez wielkości przedziałów niepewności standardowych obydwu współrzędnych x_i, y_i . W metodzie regresji ważonej (Y) proces dopasowania poszukiwanej zależności uwzględniał odcinkowy zakres zmienności tylko zmiennej y (przedziały niepewności standardowej kolejnych wartości y_i), a w metodzie zwykłej – proces ten wykorzystuje tylko ustalone pozycje punktów.

Dla danych rozważanych w tym przykładzie (tabela 4.2) obydwa upraszczające założenia o identyczności odchyleń standardowych zmiennej y oraz nieistotnym udziale niepewności zmiennej niezależnej są niespełnione. W tym przypadku zatem wybór metody regresji liniowej ważonej (X, Y) jest uzasadnionym wyborem schematu obliczeniowego. Na rysunku 4.3 widać, że podobnie jak w przypadku metody ważonej (Y), prosta regresji ważonej (X, Y) jest w znakomym stopniu dopasowana do dwóch ostatnich punktów eksperymentalnych. Punktom tym są bowiem przypisane relatywnie małe wagi.



Uwzględnienie niepewności standardowych obydwu korelowanych zmiennych w schemacie obliczeniowym regresji liniowej ma również istotne konsekwencje w procesie przewidywania wartości zmiennej niezależnej na podstawie wyznaczonej prostej. W przypadku krzywej kalibracyjnej ten aspekt regresji liniowej wiąże się z wyznaczaniem stężenia analitu w badanej próbce (x_0). Odpowiednie wyrażenie na x_0 uzyskujemy przekształcając wyznaczone równanie krzywej kalibracyjnej, a jego postać jest taka sama jak w przypadku omawiania regresji liniowej zwykłej lub ważonej (por. równania 4.12 i 4.29). W metodzie regresji ważonej (X, Y), podobnie jak w metodzie regresji ważonej (Y), konieczne jest uzyskanie serii pomiarów sygnału analitycznego (minimum 3 powtórzenia) dla badanej próbki. Umożliwia to oszacowanie standardowej niepewności wartości średniej \bar{y}_0 oraz w konsekwencji wagi $w(\bar{y}_0)$ dla tej wartości.

Analizując postać wyrażeń wyprowadzonych dla kluczowych parametrów ($a, b, u(a), u(b)$ oraz $s_{y|x,w}$) w metodzie regresji liniowej ważonej (X, Y), łatwo można zauważać analogie między ogólnymi wersjami tych wyrażeń a ich bardziej uproszczonimi wersjami wyprowadzonymi w metodzie regresji liniowej ważonej (Y) i zwykłej. Wartości x_i oraz ich wartość średnia \bar{x} w wyrażeniach metody regresji liniowej zwykłej są zastąpione przez odpowiadające im wartości oczekiwane z wagami uwzględniającymi tylko niepewności standardowe wartości zmiennej zależnej y .

(metoda regresji liniowej ważonej (Y)) lub niepewności standardowe obydwu korelowanych zmiennych x i y (metoda regresji liniowej ważonej (X,Y)). Bazując na tej analogii można zaproponować ogólne wyrażenie do wyznaczenia niepewności standardowej przewidywanej wartości zmiennej niezależnej x_0 , czyli np. stężenia odczytanego z krzywej kalibracyjnej. Korzystając z odpowiednich wyrażeń wyprowadzonych w metodzie regresji liniowej zwykłej i ważonej (Y), formułujemy ogólne wyrażenie na złożoną niepewność standardową x_0 (odchylenie standardowe x_0) [8]:

$$u(\bar{x}_0) = \frac{1}{a} \sqrt{s_{y/x,W}^2 \left(\frac{1}{w(\bar{y}_0)} + \frac{1}{\sum_{i=1}^n W_i} \right) + \frac{(\bar{y}_0 - \bar{Y})^2}{a^2} u^2(a)} \quad (4.48)$$

gdzie: $s_{y/x,W}$ jest resztowym odchyleniem standardowym regresji ważonej (X,Y) (równanie 4.47), a $w(\bar{y}_0)$ jest wagą dla pomiarów y_0 (sygnału analitycznego badanej próbki) wyrażoną w tej metodzie jako odwrotność kwadratu niepewności standarowej pomiaru \bar{y}_0 :

$$w(\bar{y}_0) = \frac{1}{u^2(\bar{y}_0)} \cong \frac{1}{s^2(\bar{y}_0)} \quad (4.49)$$

Ogólne wyrażenie na niepewność standardową przewidywanej wartości zmiennej niezależnej w postaci równania (4.48) kryje w sobie w dość złożony sposób udziały czterech źródeł niepewności, jakie towarzyszą konstrukcji liniowej zależności między zmiennymi x i y . Aby czytelnie wyeksponować te źródła niepewności, należy dokonać pewnej reorganizacji tego równania i narzucenia na potrzeby tego zadania pewnych przybliżeń. Podniesienie do kwadratu $u(x_0)$, zmiana kolejności wyrazów, wykorzystanie ogólnej definicji wag punktów W_i oraz narzucenie warunków stałości niepewności standardowych x_i oraz \bar{y}_i (tj. założenie, że $u(\bar{y}_i) = u(\bar{Y})$ oraz $u(x_i) = u(\bar{X})$ dla każdego punktu $i = 1, \dots, n$) pozwala zapisać wyrażenie na $u(x_0)$ w następującej przybliżonej postaci:

$$u^2(x_0) \cong \underbrace{\frac{(\bar{y}_0 - \bar{Y})^2}{a^4} u^2(a)}_{(1)} + \underbrace{\left(\frac{s_{y/x,W}^2}{a^2} \right) \frac{u^2(\bar{Y})}{n}}_{(2)} + \underbrace{s_{y/x,W}^2 \frac{u^2(\bar{X})}{n}}_{(3)} + \underbrace{\frac{s_{y/x,W}^2}{a^2} s^2(\bar{y}_0)}_{(4)} \quad (4.50)$$

Pierwszy człon w tym równaniu jest związany z niepewnością dopasowania punktów eksperymentalnych do założonej zależności funkcyjnej między x i y (np. krzywej kalibracyjnej), drugi – z wpływem niepewności pomiaru zmiennej zależnej y (np. pomiaru sygnału analitycznego podczas konstrukcji krzywej kalibracyjnej), trzeci – z wpływem niepewności pomiaru / wyznaczania wartości zmiennej niezależnej x (np. stężenia w przygotowanych próbkach wzorcowych do konstrukcji krzywej

kalibracyjnej), a czwarty – z precyją powtarzalności pomiaru zmiennej zależnej do przewidywania wartości zmiennej x (np. powtarzalność pomiaru sygnału analitycznego badanej próbki, dla której poszukiwanie jest stężenie analitu).

Przetestujmy teraz równanie (4.48) do wyznaczenia przedziałów dla prawdziwego stężenia analitu w badanych próbkach, które rozważaliśmy w przykładzie 4.2.

Dla \bar{y}_0 równego 4,56 i $23,40 \mu\text{A}$ (średnie z trzech pomiarów) szerokości przedziałów dla prawdziwego stężenia analitu są teraz następujące:

$x_0 \pm tu(x_0)$: $(1,80 \pm 0,35) \text{ mg/l}$ i $(10,2 \pm 2,3) \text{ mg/l}$ (w porównaniu do wyników $(1,4 \pm 1,6) \text{ mg/l}$ i $(10,8 \pm 1,4) \text{ mg/l}$ otrzymanych dla regresji liniowej zwykłej).

Parametr t odczytany z tabeli rozkładu t -Studenta dla prawdopodobieństwa 95% i $n - 2$, czyli 4 stopni swobody wynosi 2,776.

Porównanie rezultatów przewiduyań stężeń analitu w badanych próbkach metodami regresji liniowej ważonej (X, Y) i zwykłej generuje podobne wnioski jak w przypadku analizy porównawczej rezultatów przewiduyań uzyskanych metodą regresji ważonej (Y). Warto zauważyć ponownie, że wartości oznaczeń nie są od siebie drastycznie odległe, ale wielkości przedziałów ufności przewidywanych stężeń wykazują znaczące różnice. W przypadku zastosowania metody regresji ważonej (X, Y) przedział dla prawdziwego stężenia analitu jest znacznie węższy dla próbki o małym stężeniu i znacznie szerszy dla drugiej próbki. Jest to rezultat znacznie większych wag dla małych stężeń krzywej kalibracyjnej w porównaniu do dużych stężeń. W konsekwencji, niepewność standardowa oznaczeń w obszarze większych stężeń jest znaczaco większa. W tym przypadku warto też podkreślić, że wielkość niepewności standardowej przewidywanych stężeń dodatkowo uwzględnia niepewności wartości stężeń analitu w próbkach wzorcowych wziętych do konstrukcji krzywej kalibracyjnej.

▲ Przykład 4.6.

Przeprowadzono badania porównawcze dwóch metod oznaczania arsenu w próbkach wody. Analizom poddano kilkanaście wybranych próbek wody pobranych z różnych naturalnych cieków wodnych. Wyniki oznaczeń w postaci stężeń arsenianu(V) uzyskane metodą atomowej spektrometrii emisyjnej (ASE) ze zmodyfikowaną ścieżką przygotowania próbki porównano z wynikami oznaczeń uzyskanych metodą odniesienia opartą na technice atomowej spektrometrii absorpcyjnej (ASA). Dane eksperymentalne pochodzą z pracy [17]. Dla każdego oznaczenia oszacowano ponadto niepewność standardową na podstawie wielkości rozproszenia wyników pięciokrotnie powtarzanej analizy. Dane eksperymentalne w postaci: średni wynik oznaczenia \pm niepewność standardowa (odchylenie standardowe średniej) grupuje tabela poniżej.

Nr próbki	ASA [$\mu\text{g/l}$]	ASE [$\mu\text{g/l}$]
1	$8,71 \pm 1,92$	$7,35 \pm 2,07$
2	$7,01 \pm 1,56$	$7,92 \pm 2,23$
3	$3,28 \pm 0,76$	$3,40 \pm 0,96$
4	$5,60 \pm 1,26$	$5,44 \pm 1,53$
5	$1,55 \pm 0,39$	$2,07 \pm 0,59$
6	$1,75 \pm 0,43$	$2,29 \pm 0,65$
7	$0,73 \pm 0,22$	$0,66 \pm 0,19$
8	$3,66 \pm 0,84$	$3,43 \pm 0,97$
9	$0,90 \pm 0,25$	$1,25 \pm 0,36$
10	$9,39 \pm 2,07$	$6,58 \pm 1,85$
11	$4,39 \pm 1,00$	$3,31 \pm 0,93$
12	$3,69 \pm 0,84$	$2,72 \pm 0,77$
13	$0,34 \pm 0,13$	$2,32 \pm 0,66$
14	$1,94 \pm 0,47$	$1,50 \pm 0,43$
15	$2,07 \pm 0,5$	$3,50 \pm 0,99$
16	$1,38 \pm 0,36$	$1,17 \pm 0,33$
17	$1,81 \pm 0,45$	$2,31 \pm 0,66$
18	$1,27 \pm 0,33$	$1,88 \pm 0,54$
19	$0,82 \pm 0,23$	$0,44 \pm 0,13$
20	$1,88 \pm 0,46$	$1,37 \pm 0,40$
21	$5,66 \pm 1,27$	$7,04 \pm 1,98$
22	$0,00 \pm 0,06$	$0,00 \pm 0,01$
23	$0,00 \pm 0,06$	$0,49 \pm 0,15$
24	$0,40 \pm 0,15$	$1,29 \pm 0,37$
25	$0,00 \pm 0,06$	$0,37 \pm 0,12$
26	$1,98 \pm 0,48$	$2,16 \pm 0,62$
27	$10,21 \pm 2,24$	$12,53 \pm 3,51$
28	$4,64 \pm 1,05$	$3,90 \pm 1,10$
29	$5,66 \pm 1,27$	$4,66 \pm 1,31$
30	$19,25 \pm 4,18$	$15,86 \pm 4,45$

Rozwiążanie

Porównanie dwóch metod, z których jedną traktuje się jako metodę nową lub testowaną, a drugą – jako metodę odniesienia, sprowadza się do badania przesunięcia

systematycznego między wynikami pomiarów uzyskanymi tymi metodami. Jeśli wyniki pomiarów dotyczą jednego poziomu wielkości mierzonej (jedna próbka fizyczna poddana badaniom), wówczas zadanie to może realizować test istotności *t*-Studenta, porównujący wartości średnie dwóch serii pomiarowych wygenerowanych przez dwie porównywane metody pomiarowe (por. przykład 3.5). Jeśli wyniki pomiarów zgromadzono jednocześnie dla wielu poziomów wielkości mierzonej (tj. w wytyczonym zakresie zmienności mierzonej wielkości, np. dla pewnej liczby próbek fizycznych różniących się zawartością oznaczanego analitu), wówczas badania porównacze można przeprowadzić albo testem istotności *t*-Studenta parami, porównującym parami odpowiadające sobie rezultaty pomiarów badanymi metodami, albo angażując właściwy wariant metody regresji liniowej. W pierwszym przypadku mamy możliwość oceny wielkości tylko stałego błędu systematycznego. Korelacja liniowa wraz z metodą regresji liniowej daje natomiast możliwość pełnej analizy obciążenia jednej metody względem drugiej, tj. oceny istotności błędu systematycznego zarówno proporcjonalnego, jak i stałego. Tego typu analiza daje pełny obraz statystycznej zgodności wyników generowanych przez różne metody w wytycznym zakresie zmienności wielkości mierzonej.

Porównanie dwóch metod (tj. wyników pomiaru uzyskanych tymi metodami) z wykorzystaniem schematu regresji liniowej sprowadza się do trzech kroków:

- 1) dopasowanie punktów eksperymentalnych do założonej funkcji, tj. funkcji liniowej postaci $y = \alpha x + \beta$, (gdzie zmienna y reprezentuje wartości wielkości mierzonej uzyskane nową (testowaną) metodą, a zmienna x – odpowiadające im wartości wielkości mierzonej uzyskane metodą odniesienia) wykorzystując właściwy wariant metody regresji liniowej;
- 2) testowanie statystycznej istotności różnicy między uzyskaną wartością współczynnika nachylenia α a jego wartością oczekiwana, tj. 1 (czyli wartością współczynnika nachylenia badanej zależności w przypadku idealnym – całkowitym braku błędów pomiarowych);
- 3) testowanie statystycznej istotności różnicy między uzyskaną wartością współczynnika przecięcia β a jego wartością oczekiwana, tj. 0 (czyli wartością współczynnika przecięcia badanej zależności w przypadku idealnym – całkowitym braku błędów pomiarowych).

Testowanie statystycznej istotności różnic α oraz β względem odpowiednio 1 oraz 0 można przeprowadzić zgodnie z formalizmem testu *t*-Studenta porównującego eksperymentalnie wyznaczoną wartość współczynnika regresji z jego wartością oczekiwana (prawdziwą) (por. punkt 3.2.1). Alternatywnie ten sam efekt można uzyskać, analizując szerokości przedziałów ufności wyznaczonych dla α oraz β , tj. odpowiednio $\alpha \pm t(P, f)u(\alpha)$ oraz $\beta \pm t(P, f)u(\beta)$, gdzie t – wartość krytyczna rozkładu *t*-Studenta dla zadanego poziomu ufności P (standardowo 95%) i f liczby stopni swobody (dla poszukiwanej zależności liniowej $f = n - 2$, gdzie n oznacza

liczbę punktów eksperymentalnych, czyli liczbę par odpowiadających sobie wyników pomiarów). Jeśli wyznaczony przedział ufności α zawiera 1, wówczas stwierdzamy, że wyniki porównywanych metod nie są względem siebie przesunięte proporcjonalnie; testowana metoda nie generuje wyników obciążonych czynnikiem systematycznym proporcjonalnym względem wyników metody odniesienia. Jeśli wyznaczony przedział ufności β zawiera wartość oczekiwana 0, stwierdzamy ponadto, że wyniki testowanej metody nie są obciążone czynnikiem systematycznym stałym względem wyników metody odniesienia.

Postaramy się teraz te ogólne rozważania przenieść na grunt porównania dwóch metod analitycznych oznaczania As w próbkach wody. Dla porównywanych metod, tj. zaproponowanej ASE i standardowej (odniesienia) ASA, wyniki zgromadzone w szerokim zakresie wartości stężeń As oznaczonych w różnych próbkach wody. Wyników oznaczeń zgromadzonych dla danej metody nie można zatem traktować jako składników serii pomiarowej, gdyż poszczególne wyniki odnoszą się do innej próbki wody (która może charakteryzować się statystycznie inną wartością oznaczanej zawartości As). Zastosowanie tutaj schematu opartego na formalizmie testu *t*-Studenta porównującego dwie metody na podstawie zestawienia średnich dwóch serii pomiarowych lub testu parami prowadzi do zbyt ograniczonych wniosków. Najbardziej trafnym wyborem będzie zastosowanie metody regresji liniowej.

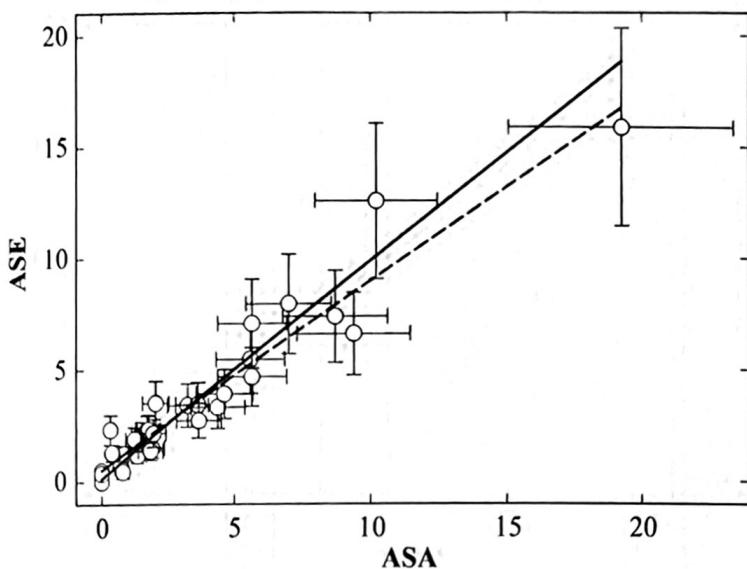
Przed przystąpieniem do obliczeń konieczne jest zweryfikowanie założeń dotyczących relacji między wartościami niepewności standardowych korelowanych zmiennych, tj. przypisanych wartościom zmiennej niezależnej (wyniki ASA) i zależnej (wyniki ASE), i ich jednorodności w wytyczonych zakresach zmienności stężenia oznaczanego analitu. Dla wyników oznaczeń metodą ASA skrajne wartości niepewności standardowych wynoszą odpowiednio 0,06 do 4,18 $\mu\text{g/l}$ i są bardzo zbliżone do skrajnych wartości niepewności standardowych wyników oznaczeń metodą ASE (odpowiednio 0,01 i 4,45 $\mu\text{g/l}$). Widać, że wartości niepewności wyników oznaczeń realizowanych metodami ASA i ASE są na bardzo zbliżonym poziomie. Dodatkowo, w ramach wyników należących do grupy oznaczeń ASA lub ASE, skrajne wartości niepewności różnią się między sobą istotnie. Wartości eksperymentalne statystyki *F* testu *F*, tj. $(4,18)^2 / (0,06)^2$ oraz $(4,45)^2 / (0,01)^2$ dla odpowiednio metody ASA i ASE zdecydowanie przekraczają wartość krytyczną tego testu nawet na poziomie ufności 99,9% ($F_{kryt}(99,9\%; 4; 4) = 53,436$).

Przeprowadzona analiza założeń regresji liniowej wskazuje jednoznacznie na konieczność zastosowania w tym przypadku ogólnego schematu obliczeń regresyjnych, czyli metody regresji liniowej (X, Y) uwzględniającej niepewności standarde obydwu korelowanych zmiennych. Poszukiwanie najbardziej prawdopodobnych wartości współczynników nachylenia i przecięcia dla zależności $y = ax + \beta$ zostanie przeprowadzone na bazie zaprezentowanej w tym rozdziale procedury iteracyjnej. Zmienna y w poszukiwanej zależności liniowej reprezentuje wyniki oznaczeń ASE, a zmienna x – odpowiadające im wyniki oznaczeń ASA. Wyniki

obliczeń w postaci wartości współczynników regresji wraz z ich przedziałami ufności na poziomie 95% zostały przedstawione poniżej w formie tabelarycznej. Rezultaty skonfrontowano z analogicznymi wynikami uzyskanymi z wykorzystaniem schematu regresji liniowej zwykłej, który ignoruje w tym przypadku nieprawność upraszczających założeń. Punkty eksperymentalne oraz proste do nich dopasowane metodą regresji liniowej ważonej (X, Y) (linia ciągła) i zwykłej (linia przerywana) ilustruje rysunek 4.4.

Regresja liniowa ważona (X, Y)	Regresja liniowa zwykła
$\alpha \pm t(95\%; 28)u(\alpha)$	
$0,973 \pm 0,183$ $\alpha = 1$	$0,8446 \pm 0,0965$ $\alpha \neq 1$
$\beta \pm t(95\%; 28)u(\beta),$	
$0,106 \pm 0,115$ $\beta = 0$	$0,544 \pm 0,526$ $\beta \neq 0$

Tabela z rezultatami obliczeń regresyjnych zawiera również wyniki testowania statystycznej istotności różnic α oraz β względem odpowiednio 1 oraz 0, przeprowadzonego na podstawie analizy szerokości ich przedziałów ufności. Jak widać, rezultaty testowania są diametralnie różne w zależności od metody regresji liniowej zastosowanej do obliczeń. Wyznaczone przedziały ufności α oraz β zawierają swoje wartości oczekiwane, tj. odpowiednio 1 i 0, jeśli obliczenia realizowane były **właściwą w tym przypadku metodą regresji ważonej (X, Y)**. Można więc postawić znak statystycznej równości między tymi współczynnikami a ich wartościami oczekiwanyimi. Obserwowane różnice między wartościami eksperymentalnymi współczynników regresji a ich oczekiwanyimi odpowiednikami mieszczą się w zakresie wielkości błędu losowego wyznaczania danych eksperymentalnych i błędu związanego z dopasowaniem punktów eksperymentalnych do założonej zależności liniowej. **Wyniki oznaczeń arsenu metodą testowaną ASE nie są więc obciążone ani czynnikiem proporcjonalnym, ani stałym względem wyników metody odniesienia ASA.** Zastosowanie niewłaściwego w tym przypadku schematu obliczeń wykorzystujących metodę regresji liniowej zwykłej doprowadziłoby do wprowadzenia całkowicie odmiennych wniosków. Wyznaczone przedziały ufności α oraz β metodą regresji zwykłej nie zawierają swoich wartości oczekiwanych. Wartości tych współczynników różnią się istotnie od swoich wartości oczekiwanych, a obserwowanych różnic nie można wyjaśnić tylko obecnością błędów losowych. Oznaczałoby to, że metoda ASE generuje wyniki oznaczeń obciążone błędem systematycznym zarówno proporcjonalnym, jak i stałym względem wyników metody odniesienia.



Rys. 4.4. Zależność wyników oznaczeń arsenu metodą ASE i ASA wyznaczona metodą regresji liniowej ważonej (X, Y) (linia ciągła) i zwykłej (linia przerywana) (punkty eksperymentalne są wprowadzone na wykres wraz z niepewnościami standardowymi wyników oznaczeń uzyskanych porównywany metodami)

Uogólniając przesłanie powyższej dyskusji, można stwierdzić, że przewidywanie wyprowadzone na podstawie metody regresji liniowej, której założenia nie są spełnione, zwykle prowadzą do błędnych wniosków. W tym przypadku zastąpienie uzasadnionej (na podstawie przeprowadzonej analizy niepewności danych eksperymentalnych) metody regresji liniowej ważonej (X, Y) jej uproszczoną wersją (metodą regresji liniowej zwykłej) doprowadziłoby do błędnej kwalifikacji metody ASE jako metody generującej wyniki obciążone względem wyników metody odniesienia ASA.

Usługa e-stat (patrz rozdział 7) zawiera moduł „Regresja liniowa” umożliwiający konstrukcję zależności liniowej metodami regresji liniowej zwykłej, ważonej (Y) i ważonej (X, Y). Moduł ten obok szczegółowej statystyki wybranej metody regresji pozwala również wyznaczyć przewidywane wartości zmiennej niezależnej wraz z niepewnością standardową na podstawie uzyskanej zależności funkcyjnej (np. przewidywanie stężenia analitu w badanej próbce wraz z przedziałem niepewności standardowej).

4.6. Regresja funkcji liniowej – podsumowanie

Zestawienie zasadniczych charakterystyk oraz zależności trzech schematów regresji liniowej zastosowanych do poszukiwania funkcji liniowej w postaci $y = ax + b$.

Regresja liniowa zwykła	Regresja liniowa ważona (Y)	Regresja liniowa ważona (X,Y)
Dane eksperymentalne		
$(x_i, y_i) \ i = 1, \dots, n$	$(x_i, \bar{y}_i, w_i) \ i = 1, \dots, n$ $w_i = w(\bar{y}_i) = \frac{1}{u^2(\bar{y}_i)} \text{ lub } w_i = \frac{n[u(\bar{y}_i)]^{-2}}{\sum_{i=1}^n [u(\bar{y}_i)]^{-2}}$	$(x_i, \bar{y}_i, W_i) \ i = 1, \dots, n$ $W_i = \frac{1}{u^2(\bar{y}_i) + a^2 u^2(x_i)}$

Charakterystyka statystyczna zależności liniowej

Współczynnik nachylenia oraz niepewność standardowa współczynnika nachylenia

$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$	$a = \frac{\sum_{i=1}^n w_i (x_i - \bar{x}_w)(\bar{y}_i - \bar{y}_w)}{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}$	$a = \frac{\sum_{i=1}^n W_i \lambda_i (\bar{y}_i - \bar{y}_w)}{\sum_{i=1}^n W_i \lambda_i (x_i - \bar{x}_w)} (*)$
$u(a) = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$	$u(a) = \frac{s_{y/x,w}}{\sqrt{\sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}}$	$u(a) = \frac{s_{y/x,W}}{\sqrt{\sum_{i=1}^n W_i (x_i - \bar{x})^2}}$ $\lambda_i = W_i \left[\frac{(x_i - \bar{x}_w)}{w(\bar{y}_i)} + \frac{a(\bar{y}_i - \bar{y}_w)}{w(x_i)} \right]$ $\bar{X}_i = \bar{x}_w + \lambda_i$

Regresja liniowa zwykła	Regresja liniowa ważona (Y)	Regresja liniowa ważona (X,Y)
Charakterystyka statystyczna zależności liniowej		
Współczynnik przecięcia oraz niepewność standardowa współczynnika przecięcia		
$b = \bar{y} - a\bar{x}$ $u(b) = s_{y/x} \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}$	$b = \bar{y}_w - a\bar{x}_w$ $u(b) = s_{y/x,w} \sqrt{\frac{\sum_{i=1}^n w_i x_i^2}{n \sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}}$	$b = \bar{y}_w - a\bar{x}_w$ $u(b) = s_{y/x,w} \sqrt{\frac{1}{\sum_{i=1}^n W_i} + \frac{\bar{X}^2}{\sum_{i=1}^n W_i (\bar{X}_i - \bar{X})^2}}$
Resztowe odchylenie standardowe		
$s_{y/x} = \sqrt{\frac{\sum_{i=1}^n (y_i - ax_i - b)^2}{n-2}}$	$s_{y/x,w} = \sqrt{\frac{\sum_{i=1}^n w_i (\bar{y}_i - ax_i - b)^2}{n-2}}$	$s_{y/x,w} = \sqrt{\frac{\sum_{i=1}^n W_i (\bar{y}_i - ax_i - b)^2}{n-2}}$
Przewidywanie wartości zmiennej niezależnej		
$\bar{x}_0 = \frac{\bar{y}_0 - b}{a}$ $u(\bar{x}_0) = \frac{s_{y/x}}{a} \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(\bar{y}_0 - \bar{y})^2}{a^2 \sum_{i=1}^n (x_i - \bar{x})^2}}$	$\bar{x}_0 = \frac{\bar{y}_0 - b}{a}$ $u(\bar{x}_0) = \frac{s_{y/x,w}}{a} \sqrt{\frac{1}{w_0} + \frac{1}{n} + \frac{(\bar{y}_0 - \bar{y}_w)^2}{a^2 \sum_{i=1}^n w_i (x_i - \bar{x}_w)^2}}$ $w_0 = w(\bar{y}_0) = \frac{n [u(\bar{y}_0)]^{-2}}{\sum_{i=1}^n [u(\bar{y}_i)]^{-2}}$	$\bar{x}_0 = \frac{\bar{y}_0 - b}{a}$ $u(\bar{x}_0) = \frac{1}{a} \sqrt{s_{y/x,w}^2 \left(\frac{1}{w(\bar{y}_0)} + \frac{1}{\sum_{i=1}^n W_i} \right) + \frac{(\bar{y}_0 - \bar{Y})^2}{a^2} u^2(a)}$ $w(\bar{y}_0) = \frac{1}{u^2(\bar{y}_0)}$

(*) wyznaczanie metodą iteracyjną

4.7. Linearyzacja funkcji

Rozszerzenie zakresu roboczego metody lub natura badanego procesu/zjawiska są często powodem pojawienia się odstępstw od liniowej zależności między sygnałem analitycznym a stężeniem analitu lub inną zmienną niezależną. Jeśli znany jest model teoretyczny (postać funkcyjna), który może opisywać z dużym prawdopodobieństwem obserwowane zmiany, to w pewnych przypadkach można dokonać estymacji parametrów obserwowanej zależności nieliniowej metodą regresji liniowej. Ułatwieniem będzie znalezienie sposobu przekształcenia przewidywanej funkcji nieliniowej do postaci liniowej. Taka transformacja nosi nazwę **linearyzacji funkcji**. Współczynniki funkcji zlinearyzowanej wyznacza się metodą regresji liniowej, a następnie równanie funkcji zlinearyzowanej przekształca się z powrotem do postaci funkcji pierwotnej. Przykłady linearyzacji kilku typowych funkcji nieliniowych zawiera tabela 4.3.

Tabela 4.3. Typowe funkcje (modele) nieliniowe oraz ich zlinearyzowane odpowiedniki

Funkcja pierwotna (nieliniowa)	Funkcja zlinearyzowana
$y = Bx^A$	$\ln \underline{y} = \frac{\ln B}{b} + \frac{A}{a} \ln \underline{x}$
$y = B \cdot A^x$	$\ln \underline{y} = \frac{\ln B}{b} + \frac{\ln A}{a} \underline{x}$
$y = B \cdot e^{Ax}$	$\ln \underline{y} = \frac{\ln B}{b} + \frac{A}{a} \underline{x}$
$y = B + A \ln x$	$\underline{y} = \frac{B}{b} + \frac{A}{a} \ln \underline{x}$
$y = (B + Ax)^2$	$\sqrt{\underline{y}} = \frac{B}{b} + \frac{A}{a} \underline{x}$
$y = \frac{Bx}{A+x}$	$\frac{1}{\underline{y}} = \frac{1}{B} + \frac{A}{B} \cdot \frac{1}{\underline{x}}$

▲ Przykład 4.7.

Wałąną cechą preparatów radioizotopowych stosowanych w diagnostice medycznej jest szybki zanik aktywności promieniotwórczej radioizotopu wprowadzonego do organizmu. Własność tę można scharakteryzować śledząc zmiany aktywności promieniotwórczej z czasem. Badaniom takim poddano nową formułę preparatu zawierającego radioaktywny izotop technetu (^{99}Tc). Wykonano pomiary względnego natężenia

promieniowania świeżo sporzązonego preparatu w funkcji czasu. Uzyskano następujące wyniki:

t [godz.]	0	1	3	5	7	9
$\frac{N(t)}{N_0}$	1,000	0,891	0,708	0,562	0,447	0,355

Po jakim czasie aktywność promieniotwórcza izotopu ^{99}Tc spadnie do połowy swojej początkowej wartości?

Rozwiązanie

Rozpad promieniotwórczy opisuje model wykładniczy

$$\frac{N(t)}{N_0} = B \cdot e^{-kt}$$

gdzie: $N(t)$ i N_0 oznaczają odpowiednio aktywność (liczbę rozpadów) izotopu promieniotwórczego w czasie t i w chwili początkowej ($t = 0$), k jest stałą szybkości rozpadu, a B jest stałą (teoretycznie równą 1).

Logarytmując obustronnie powyższe równanie otrzymujemy zlinearyzowaną postać prawa rozpadu promieniotwórczego, tj.

$$\ln\left(\frac{N(t)}{N_0}\right) = \ln B - kt$$

Uzyskana funkcja wskazuje na liniową zależność między wartościami $\ln\left(\frac{N(t)}{N_0}\right)$ a wartościami zmiennej t . Formalnie można ją zapisać jako $y' = b + ax'$, przyjmując następujące podstawienia:

$$y' = \ln\left(\frac{N(t)}{N_0}\right)$$

$$x' = t$$

$$a = -k$$

$$b = \ln B$$

Wartości korelowanych zmiennych zestawione są poniżej.

t [godz.]	0	1	3	5	7	9
$\ln\left(\frac{N(t)}{N_0}\right)$	0,00000	-0,11541	-0,34531	-0,57625	-0,80520	-1,03564

Aby uniknąć błędów zaokrągleń, szczególnie dotkliwych w przypadku prze kształceń wykładniczych, zaleca się zwiększenie liczby cyfr znaczących w wynikach pośrednich.

Współczynnika kierunkowego (nachylenia) a i współczynnika przecięcia b z linearzowanej zależności oraz ich odchyлеń (niepewności) standardowych (s_a i s_b) poszukiwać będziemy metodą regresji liniowej zwykłej, korzystając z wzorów (4.3), (4.4), (4.8) oraz (4.9). Równanie wyznaczonej prostej ma następującą postać:

$$y' = -0,00026 - 0,115x'; r = 0,999999$$

gdzie: $a = -0,11505/\text{godz.}$ i $s_a = 5,76 \cdot 10^{-5}/\text{godz.}$ oraz $b = -0,00026$ i $s_b = 0,000302$

Kolejnym etapem jest przekształcenie równania wyznaczonej zlinearzowanej zależności z powrotem do pierwotnej postaci nieliniowej. Zadanie sprowadza się więc do wyznaczenia współczynników k i B na podstawie obliczonych współczynników regresji liniowej a i b

$$a = -k, \text{czyli } k = 0,11505/\text{godz.}$$

$$b = \ln B, \text{czyli } B = e^b = 0,9997$$

Odchylenia standardowe współczynników funkcji nieliniowej wyznaczamy, stosując reguły propagacji niepewności na podstawie równania (2.17)

$$s_k = s_a = 5,8 \cdot 10^{-5}/\text{godz.}$$

$$s_B = \left| \frac{d}{db} B \right| s_b = \frac{d}{db} e^b s_b = e^b s_b = 0,00030$$

Równanie pierwotnej funkcji nieliniowej przedstawić można zatem jako

$$\frac{N(t)}{N_0} = 0,9997 \cdot e^{-0,11505t}$$

Jeżeli znamy charakterystykę statystyczną zmian aktywności promieniotwórczej testowanego preparatu, jesteśmy w stanie określić czas połowicznego zaniku aktywności promieniotwórczej izotopu ^{99}Tc . Poszukujemy zatem takiej wartości zmiennej niezależnej (czasu), dla której $\frac{N(t)}{N_0} = 0,5$, czyli $\ln\left(\frac{N(t)}{N_0}\right) = -0,693$. Czas ten oznaczymy symbolem τ i wyznaczmy na podstawie obliczonej zależności liniowej, korzystając z równania (4.12). Wartość τ oraz jego odchylenie standardowe, s_τ oszacowane na podstawie równania (4.16), wynoszą odpowiednio 6,0212 oraz 0,0043 godz. Przedział ufności dla wyznaczonego czasu połówkowego i dla poziomu ufności 95% wynosi zatem

$$\tau \pm ts_\tau = 6,021 \pm 0,012 \text{ godz.}$$

Parametr t jest współczynnikiem rozszerzenia niepewności standardowej τ , odczytanym z tabeli rozkładu t -Studenta dla poziomu ufności 95% i 4 (liczba punktów eksperymentalnych pomniejszona o 2) stopni swobody.

Linearyzacja pozwala stosunkowo łatwo wyznaczyć parametry zależności nieliniowej. Należy jednak pamiętać, że ogranicza się ona tylko do niektórych modeli funkcji nieliniowych (por. tabela 4.3) oraz bardzo często nie spełnia założeń regresji liniowej. Regresja liniowa zakłada przede wszystkim gaussowski charakter rozrzutu punktów eksperymentalnych wokół przewidywanej krzywej. Linearyzacja zaś w wyniku zastosowania pewnej transformacji „zniekształca” ten rozrzut, często powodując nawet jego zwiększenie. Efektem linearyzacji jest również znaczne zróżnicowanie niepewności standardowych poszczególnych punktów. W konsekwencji współczynniki regresji zlinearyzowanej funkcji wyznaczone metodą regresji liniowej będą charakteryzować się mniejszą dokładnością niż ich odpowiedniki obliczone metodą regresji nieliniowej. Ta ostatnia, choć bardziej złożona matematycznie, jest uniwersalną metodą analizowania zależności wśród danych eksperymentalnych, wolną od wspomnianych wyżej ograniczeń.